图书情报工作
第63卷 第24期 2019年12月

社会化标注系统标签质量影响因素研究：基于随机森林算法*

■ 张云中 秦艺源
上海大学图书情报档案系 上海 200444

摘要：［**目的/意义**］在社会化标注系统中，标签质量往往关乎用户对网络资源的分类、查询、浏览、获取等使用体验，确定影响标签质量的关键因素有助于进一步优化社会化标注系统的资源组织核心功能。［**方法/过程**］以社会化标注系统的标签为研究对象，从标注主体、标注客体、标注环境、标注动机、标注方式、标注产物等维度入手重构标签质量影响因素模型，尝试探究影响社会化标签质量的关键因素，并运用问卷调查方法收集数据，结合有监督学习的随机森林算法，建立标签质量影响因素的决策树模型。［**结果/结论**］结果显示，标注主体是影响标签质量的首要关键维度，主体的知识结构和认知水平、标注频度及其感知有用性对标签质量的影响突出；标注方式是影响标签质量的次要关键维度，标签推荐和规范标签提示是影响标签质量的重要因素。

关键词：社会化标注系统 标签质量 机器学习 随机森林
分类号：G250 TP181
DOI：10.13266/j.issn.0252-3116.2019.24.013

1 引言

社会化标注系统（Social Tagging System，STS）作为Web2.0时代新兴的网络信息资源组织和管理系统，允许用户自由地对网络资源进行描述和标注，产生的标签对于网络资源的组织和索引非常有效[1]。STS是Web2.0网络环境下用户依据自我认知对网络资源自由地张贴标签以实现网络资源描述、分类及导航的平台[2]。当然，用户所张贴标签的质量往往有优劣之分。标签质量是对标签能否精准地描述待标注的网络资源，以方便用户对网络资源的分类、查询、浏览、获取及利用的精准程度的衡量，其往往与用户、网络资源、标注环境、标注动机、标注方式及标注产物等因素密切相关。当前社会化标注系统中的标签数量急剧增多，但质量却参差不齐，这极大地影响了用户对网络信息资源的分类、查询、浏览、获取及分享等体验。就此问题，国内外众多学者从标签评估和推荐出发，通过对标签质量评估来推荐高质量的标签供用户使用，而对于如何减少低质量标签的出现、及出现的原因却研究甚少[3]。标签质量既有高低之分，便有影响其高低的因素，深入研究影响标签质量的各个因素，锁定影响标签质量的关键因素是从根本上提高标签质量的关键。因此，笔者从影响标签质量的各个因素出发，计算各个影响因素对于标签质量的影响权重，这不仅能从根本上解决标签质量评估权重确定的问题，采用权重系数与现有统计指标体系相结合可以更好地完善标签质量评估模型，而且社会化标注系统也可据此采取有效的改进措施来提高标签质量，促进网络信息资源的组织和索引，这对Web2.0环境下以用户为中心的社会化标注系统的功能完善具有重要参考意义。

2 文献回顾

目前，多数学者并不将标签质量影响因素作为独立的研究议题，而是将其作为标签质量评估研究的一个环节。就秉持的观点而言，也有多维度综合影响因素及单维度关键影响因素两种类别：

（1）多维度综合影响因素。此类研究认为标签质量受标注主体、标注环境、标注动机以及标注方式等综合因素共同作用的影响，章成志团队即是其中代表，其

* 本文系国家哲学社会科学基金项目“基于形式概念分析的社会化标注系统语义发现与语义映射研究”（项目编号：16CTQ023）研究成果之一。
作者简介：张云中（ORCID：0000-0002-7323-2561），副教授，博士，硕士生导师，E-mail：zhang-yun-zhong@126.com；秦艺源（ORCID：0000-0003-4101-8584），硕士研究生。
收稿日期：2019-04-28 修回日期：2019-07-22 本文起止页码：119-126 本文责任编辑：徐健

主要观点是标注主体过于自由随意、过于主观是导致标签质量低下的重要原因,标注系统功能机制的不健全等标注环境因素也会影响标签质量,规范标签提示、标签拼写提示等标注方式因素可以减少标签错拼、歧义和同义的现象,同时标注动机的不同会导致标签质量的差异[4-5]。

(2)单维度关键影响因素。此类研究倾向于标签质量受某个关键因素的影响,其中认为"标注方式"是关键影响因素的居多。黄如花在对 WorldCat、Flickr、Bibsonomy 以及豆瓣的标签质量控制进行研究时指出,可从规范标签提示、标签拼写提示、检错机制、以及标注指南等方面提高标签质量[6];吴方枝在对 Flickr 的标签质量控制研究中也持相似见解,并补充指出重视热门标签管理也可提高标签质量[7];N. Sogol 等提出通过标签推荐可提高标签质量[8];C. Hall 等提出可引用受控词表进行标签推荐进而达到提高标签质量的目的[9];M. Guy 也强调输入提示、拼写检查、标签推荐等标注方式对标签质量有着重要影响[10]。

当然,也有些学者认为"标注环境"是关键因素,特别是社会化标注系统的界面会对标签质量产生影响。F. Floeck 等指出社会化标注系统的界面设计会影响标签的质量,强调了标注环境对于标签质量的影响[11];S. Sen 等认为改进社会化标注系统界面可以达到提高标签质量的目的[12]。

更有部分学者强调"标注主体"才是关键因素,朱庆华认为用户规模、用户结构以及用户的标注频度等主体特征会对标签检索质量产生影响[13];罗琳通过实证研究得出豆瓣图书标签的信息质量、感知有用性、感知易用性会正向影响主体的标注意愿[14]。

无论是秉持多维度综合影响因素的观点,还是单维度关键影响因素的看法,学者们就社会化标注系统标签质量的可能影响因素已能求同存异,但遗憾的是当前研究多侧重于关注有哪些相关因素,而未系统地探寻这些因素对于标签质量影响的一般性权重关系。这恰恰是本研究所要尝试解决的问题。

3 研究方法

承上而言,本研究旨在探寻多影响因素对标签质量影响的权重,本质上是属于权重确定问题,最常用的定权方法有主成分分析法、多元回归分析、层次分析法及基于支持向量机(SVM)的机器学习方法等。

主成分分析法适用于多变量转化为少数几个综合特征(主成分)的降维问题中的权重确定[15];多元回归分析通过显著性水平来衡量因素对变量的影响,但回归方程假设严格,需知引起因变量改变的所有解释变量的因素,否则易出现伪回归问题;层次分析法适用于多方案择优,且在指标过多时会出现判断矩阵阶数变大,赋值困难,精度较差等问题;SVM 机器学习算法虽理论完善,但其多适用于二分类问题,对于多分类问题效果较差。

综合考量下,本研究拟选用随机森林的机器学习算法建立影响因素特征与标签质量之间的预测模型,进而用分类器加以分类预测。随机森林在处理该问题时具有以下优势:①相对于主成分分析法,随机森林能够通过对数据资料的客观训练直接得到各个影响因素的权重,而前者必须将众多影响因素通过降维进而得到主成分的权重,且各个影响因素的权重无从得知;②相对于多元回归分析,随机森林不需要穷尽引起因变量改变的所有解释因素,而且随机森林具有样本随机性和特征随机性,抗拟合能力强,可避免伪回归问题;③相对于层次分析法,随机森林能分析大规模样本,抗噪声能力强,得到的影响因子更可靠,精度更高[16];④相对于 SVM 等机器学习模型,随机森林得到的决策树模型容易解释,以 if-then 的规则形式建立的影响因素特征和标签质量之间的关系,通俗易懂、易于理解和应用。

4 标签质量影响因素模型

笔者在相关研究提出的标签质量影响因素基础上,通过阅读大量国内外相关文献,访谈专业研究人员以及社会化标注系统用户等方式,最终提出标签质量影响因素模型,该模型从标注主体、标注客体、标注环境、标注动机、标注方式以及标注产物 6 个维度度量各因素对标签质量的影响。见图 1。

(1)标注主体即标注者,其学科背景、知识结构和认知水平、标注频度、兴趣偏好以及标注情绪等因素必会对标签质量产生影响;

(2)标注客体即标注对象,也就是待标注的网络资源,其数量、质量和类型对于标签质量也会产生一定的影响;

(3)标注环境主要指用户实施标注行为的各种社会化标注系统平台,其功能是否完整、性能是否优越以及平台是否稳定均会影响标签质量;

(4)标注动机主要指用户产生标注行为的动力和原因。标注动机一般涵盖揭示资源主题/分类/属性(关于什么)、描述资源载体/类型(是什么)、拥有者、修饰标签(细化和限定现有的标签)、描述资源的特征、自我引用与参考、任务与个人资源管理 7 种[17];

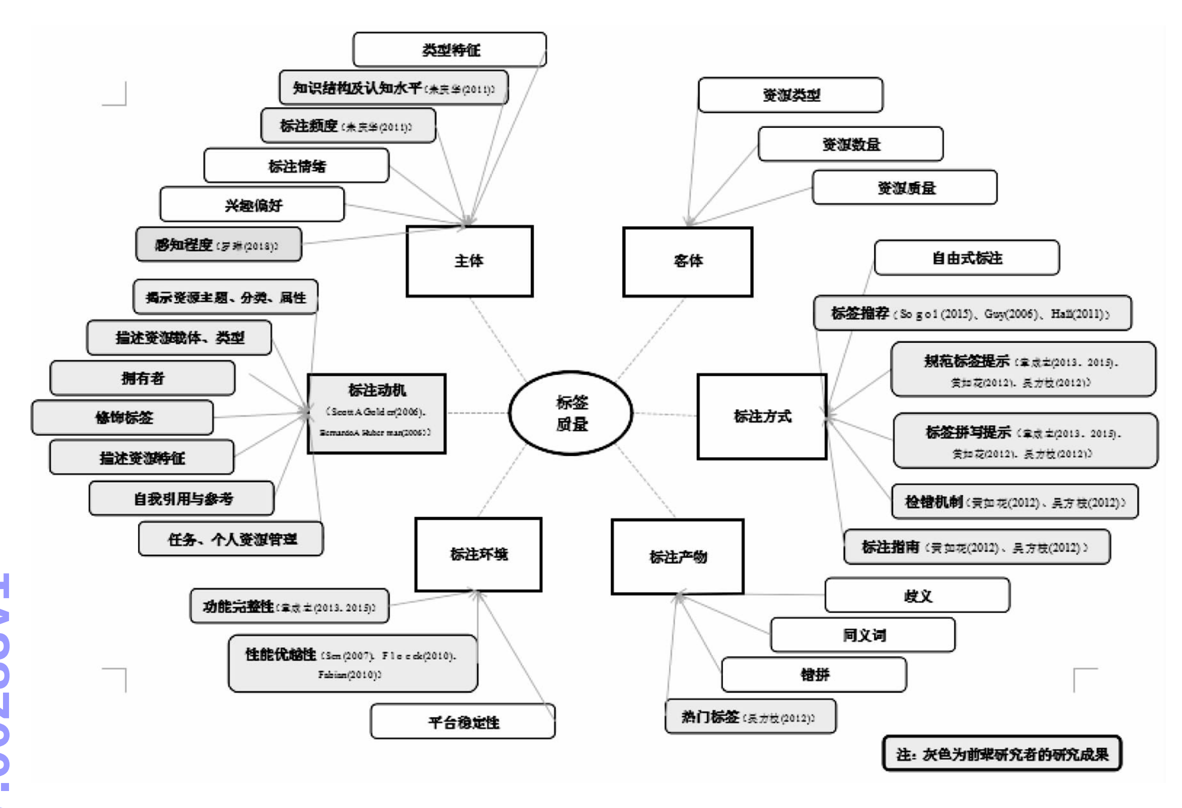


图1　标签质量影响因素模型

（5）标注方式主要包括自由式标注和干预式标注。标签推荐、规范标签提示、标签拼写提示、标注指南、检错机制是目前主要的5种干预式标注机制；

（6）标注产物即标签本身。标签的质量与自身形态有直接关系，歧义、同义、错拼等形态的标签占比高是标签质量低的直接表现。

5　社会化标签质量影响因素实证研究

5.1　问卷设计

根据笔者提出的模型假设以及问卷设计的一般规律，将问卷设计为3个板块，见表1。

（1）基本信息板块。虽然这些基本信息不直接属于笔者提出的标签质量影响因素模型，但是考虑到其对预测结果可能会产生影响，故将其加入问卷中，这一板块主要包含Q1－Q6共6个问题。

（2）影响因素特征板块。这一板块主要包括Q7－Q35共29个关于标签质量影响因素的问题，根据李克特量表的法则，涵盖标注主体、标注客体、标注环境、标注动机、标注方式及标注产物6大主要维度。用户将按照对其接受程度（非常不同意、不同意、一般、同意、非常同意）对每个问题进行选择。

（3）目标特征板块，这一板块包括Q36一个问题，表征了标签质量的5个等级，用户通过填写对这个说法的接受程度，来量化其标签质量的高低。

5.2　数据采集及检验

为确保研究结果的科学性和准确性，本研究将问卷发放对象划分为具有信息组织背景的用户和普通用户。发放渠道主要有两种（括号内为收回问卷数量）：①通过E-Mail向从事信息组织特别是社会化标注系统研究的高校与科研机构的教师（76人）及学生（113人）发放电子问卷；②通过分享问卷链接和二维码向豆瓣（图书（18人）、电影（27人）、音乐（31人））、Flickr（17人）、博客（44人）、Diigo（19人）、Pinterest（28人）、好网角（19人）等多个社会化标注平台用户在线发送问卷。2个月共发放问卷523份，收集调查问卷429份，其中有效问卷392份，问卷回收率为82%。在样本中，男女性别比为136:256；年龄大多分布在21－30岁，共277人，这也是社会化标注平台的主要用户人群，21岁以下36人，31－40岁53人，40岁以上26人；调查对象的文化程度大多集中在本科学士学位（137人）和硕士学位（172人），专科、高中及以下（学士学位以下）和博士学位较少，分别为36人和47人；其中普通用户、具有信息组织学科背景专业用户的比例为203:189，接近1:1；在社会化标注系统中发布过网络资源的人数与未发布过的人数之比为164:228；接触任意社会化标注系统（如豆瓣网）的人数占86%，说明绝大

表 1 问卷设计表

维度	编号	问项及选项
		基本信息板块
	Q1	性别(男/女)
	Q2	年龄(21 岁以下,21 -30 岁,31 -40 岁,40 岁以上)
	Q3	文化程度(专科、高中及以下(学士学位以下),本科(学士学位),硕士学位,博士学位)
	Q4	背景(普通人员/信息组织学科背景的专业人员)
	Q5	网络资源拥有者及发布者(在社会化标注系统中发布过网络资源)(是/否)
	Q6	接触任意社会化标注系统(如豆瓣网)的时间(从不,1 年以内,1 年 -3 年,3 年以上)
		影响因素特征板块
标注主体	Q7	掌握网络资源分类专业知识的用户较之未受该类训练者更能标注高质量的标签
	Q8	用户对网络资源内容及相关领域的知识结构和知识背景越熟悉,标注的标签质量越高
	Q9	用户对社会化标注系统的标注功能越熟悉,标注的标签质量越高
	Q10	用户标注网络资源时的心情状态越好,标注的标签质量越高
	Q11	用户对网络资源内容的兴趣偏好越强,标注的标签质量越高
	Q12	当用户认为社会化标注流程容易学习和操作时,标注的标签质量更高
	Q13	当用户意识到标签有助于自身及他人对网络资源的检索、推荐和分享时,标注的标签质量更高
标注客体	Q14	相对于富媒体类型网络资源(如图片、视频、音乐等),用户对以文本类型为主的网络资源添加的标签质量更高
	Q15	社会化标注系统中,待标注网络资源的数量越多,用户标注的标签质量越高
	Q16	社会化标注系统中,待标注网络资源的质量越高,用户标注的标签质量越高
标注环境	Q17	用户在功能完整(涵盖标注、检索、导航、推荐、群组、分类等功能)的社会化标注系统中标注的标签质量更高
	Q18	用户在性能优越(响应速度快、界面友好、容易操作等)的社会化标注系统中标注的标签质量更高
	Q19	用户在平台稳定性强(兼容性好、容错率高、维护及时、资源更新及时、商业运营稳定等)的社会化标注系统中标注的标签质量更高
标注动机	Q20	用以揭示资源主题标注的标签较之其它标签质量更高
	Q21	用以揭示网络资源载体或类型的标签(如书籍、文章、博客)较之其它标签质量更高
	Q22	用以揭示网络资源的作者、协作者的标签较之其它标签质量更高
	Q23	用以修饰现有标签的标签较之其它标签质量更高
	Q24	用以描述资源特征的标签(如有趣、雷人)较之其它标签质量更高
	Q25	用以个人参考(如我买过的,我的书)的标签较之其它标签质量更高
	Q26	用以自身任务管理(如找工作)的标签较之其它标签质量更高
标注方式	Q27	提供标签推荐机制(如热门标签等非受控推荐或叙词表、主题词表等受控词表推荐)的标注方式较之不加任何干预的标注方式,前者标注的标签质量更高
	Q28	提供标签规范提示(用户输入标签时,系统将可能匹配的规范标签自动推荐给用户)的标注方式较之不加任何干预的标注方式,前者标注的标签质量更高
	Q29	提供标签拼写提示(如符号限制、标签长度限制)的标注方式较之不加任何干预的标注方式,前者标注的标签质量更高
	Q30	提供标签检错机制(大小写、单复数、简缩写、结合词以及简繁词等词汇控制、输入法纠错等)的标注方式较之不加任何干预的标注方式,前者标注的标签质量更高
	Q31	提供标注指南(如"什么是标签"、"标注入门指导")的标注方式较之不加任何干预的标注方式,前者标注的标签质量更高
标注产物	Q32	不存在歧义的标签质量更高
	Q33	不存在同义词的标签质量更高
	Q34	不存在错拼现象的标签质量更高
	Q35	热门标签较之其它标签的质量更高
		目标特征板块
	Q36	社会化标注系统中,"标签质量"对于衡量"标签能否精准地描述待标注的网络资源,以方便用户对网络资源的分类、查询、浏览、获取及利用"至关重要

部分调查对象对社会化标注系统的功能和标注过程均有一定了解,其中接触时间 1 年以内的有 106 人,1 年 -3 年的有 74 人,3 年以上的 157 人。

笔者之后对 392 份问卷所得的数据导入 SPSS 21.0 进行信效度检验。本研究问卷量表的信度检验方法采用常用的 Cronbach's Alpha 系数。结果显示,Alpha 的值为 0.931 >0.8,说明量表信度很好;问卷量表的结构效度采用 KMO 和 Bartlett 进行检验。结果效度系数 KMO 的值为 0.912 >0.7,Bartlett 的球形度检验卡方值为5 661.566,df =435,sig =0.000,具有统计学意义,说

明问卷量表的结构设计合理。

5.3 数据预处理

通过问卷调查收集到的数据都是以文本形式存在的，不能直接用于机器学习的模型训练和测试过程，因此首先要对其进行量化处理。

5.3.1 有序特征的量化　对于 Q2、Q7 - Q36 等所有利用李克特量表进行统计的问题，5 个选项之间都存在明显的顺序关系，各个选项之间的差异反映用户的接受程度，这类特征被称为有序特征，在进行量化时可直接将各个选项按顺序赋予一个整数值。如 Q8 在测量主体的知识结构和认知水平对标签质量的影响时，5 个选项按照用户对其接受程度进行顺序排列，其值可以用 1 - 5 这 5 个自然数表示。因此 Q7 - Q36 特征均被量化到 1 - 5 之间，Q2 特征的取值在 1 - 4 之间。由于机器学习对特征间的尺度差异很敏感，因此还需采用Z-score 标准化方法对特征进行归一化处理[18]，利用转化函数将特征取值不同的 Q2 与 Q7 - Q36 统一到相同的取值空间，转化函数如公式(1)所示：

$$X^* = \frac{x-\mu}{\sigma} \qquad 公式(1)$$

其中，X 为未归一化的单个特征，μ 为所有未归一化特征的平均值，σ 为未归一化特征的标准差，X^* 为归一化后的特征。归一化的特征为 0 - 1 之间的实数，保留着特征中存在的有序信息。

5.3.2 无序特征的量化　对于 Q1 性别、Q3 文化程度、Q4 背景、Q5 是否是信息资源拥有者及发布者、Q6 接触 STS 的时间等问题，各个选项之间不存在明显的数据差异，为无序特征。对于无序特征的量化一般采用独热编码方式，即将十进制编码转化为稀疏表示的多位二进制编码，生成的二进制编码中，只有一位为 1，其它位均为 0。以 Q1 为例，独热编码分别为：男 1 0，女 0 1。

5.4 模型训练与评估

笔者选择随机森林算法建立决策树模型，为提高模型的精度和预测准确率，将收集到的 392 份问卷作为样本进行模型训练，并采用留一法验证模型精度，模型误差为 0.1475，预测准确率（即模型精度）达 85.25%。每一个样本中均包括 Q1 - Q35 这 35 个影响因素特征和 Q36 这 1 个待预测的标签质量，其中标签质量由李克特量表的五级评分表示，因此，建模本质是一个五分类的有监督学习问题，分类目的在于得出各个影响因素特征对标签质量的影响权重。

5.4.1 信息增益和影响因子　利用随机森林算法测量各个影响因素特征对标签质量的影响权重，其关键指标在于信息增益。在决策树学习算法中，信息增益是特征选择的一个重要指标，用以衡量一个特征能够为分类特征贡献信息的大小，如其贡献的信息越多，说明该特征越重要，相应的信息增益就越大。本研究中信息增益用以表现各个特征对于数据集的重要程度。在计算完各个影响因素特征的信息增益后，会根据信息增益的高低对样本数据进行分类排序。在决策树建立时，也会根据信息增益的高低对样本数据进行划分直至划分至叶子结点，如此可得到各个影响因素的相对重要性排序。因此，信息增益可理解为各个影响因素特征的影响因子，其具体计算方法如下：给定样本 D 和特征 a，依据与特征 a 相对应的决策准则可以将样本 D 分成 n 个子集 D1，D2，…，Dn。则特征 a 对于样本 D 的信息增益如下，H(D) 和 H(D|a) 分别指样本 D 的信息熵以及样本 D 关于特征 a 的交叉信息熵。

$$g(D|a) = H(D) - H(D|a_ \qquad 公式(2)$$

$$H(D) = -\sum_{k=1}^{k}\frac{ck}{|D|}\log\frac{ck}{|D|} \qquad 公式(3)$$

$$H(D|a) = -\sum_{i=1}^{n}\frac{|Di|}{|D|}H(Di) \qquad 公式(4)$$

结合公式(2)、(3)、(4)可以看出，信息增益代表引入特征 a 之后，样本 D 不确定性的降低程度，因此信息增益可以用来反映特征 a 对样本 D 产生的影响。

5.4.2 阈值　按信息增益对各个影响因素特征进行排序后，随机森林中决策树的分类规则就大致确定了。除此之外，还需对决策节点的影响因素特征选择阈值。阈值即决策树的分类规则，合理的阈值选择能使样本被尽可能准确地划分至其应属的类别。

关于阈值优确定的方法有很多，对于离散特征，最简单的是先穷举特征可能取到的所有值，把所有值都作为阈值，计算各阈值下特征的信息增益水平，然后按需选择其中一个或多个作为真正的阈值，以使得在该阈值的作用下特征能对数据集产生最大的信息增益。在本研究中，Q8 是对标签质量影响最大的特征，排在第 1 位，其阈值为 4.53。在对一个样本进行测试时，首先判断其特征 Q8 的值(6)是否大于 4.53，从而将其划入左侧决策结点。再判断特征 Q27 的值(2)是否大于 3.82，从而将其划入右侧决策节点……依此类推。

5.4.3 模型训练与评估　影响因素特征排序和阈值确定之后，即可建立随机森林的各个决策树，进而得到集成算法的最终结果，完成标签质量影响因素的决策树模型的构建。

模型建立之后，采用留一法对其进行评估及测试，

本研究共有 392 份样本数据，每次随机抽取 1 个样本作为测试集，剩下的 391 份样本作为训练集，总共随机抽取了 100 次，结果指出 100 个测试数据的正确率达 92%，即模型预测准确率达到 92%，故模型具有较好的预测效果。

图 2 展示了其中一个样本的预测过程。决策树示意图中矩形表示决策结点，圆圈表示机会结点，三角形表示叶子结点。模型训练时计算出特征 Q8 的信息增益最大，因此选择其作为第 1 次决策的判断依据，而最佳阈值为4.53。当输入测试样本的 Q1 – Q35 这 35 个特征值之后，首先判断其特征 Q8 的值（6）是否大于 4.53，从而将其划入左侧的决策结点；接下来将 Q27 作为第 2 次决策的依据，其阈值为 3.82，从而划入右侧的决策节点。接下来依次以信息增益次高的特征为标准，重复上述过程，直至达到叶子结点，从而得到决策：该样本的标签质量等级即 Q36 为 3。图 2 中加粗的箭头表示了这一决策的过程。对于每一个测试样本，只要给定其特征的值（对调查问卷中前 35 个问题的答案），均可以按这个方法推测出标签质量的等级（Q36）。

5.4.4 影响因子分析 在模型训练和评估过程中，随机森林算法即可根据样本数据计算出各个影响因素特征的信息增益，即各个影响因素特征的影响因子（见图 3）。图像横轴以调查问卷中问题的编号来标记各个特征，而纵轴则以信息增益反映各特征的影响因子。进一步，可通过 PCA 降维算法计算出各个维度的影响因子（见图 4）。图像横轴以维度名称来标记，纵轴则以信息增益反映各维度的影响因子。

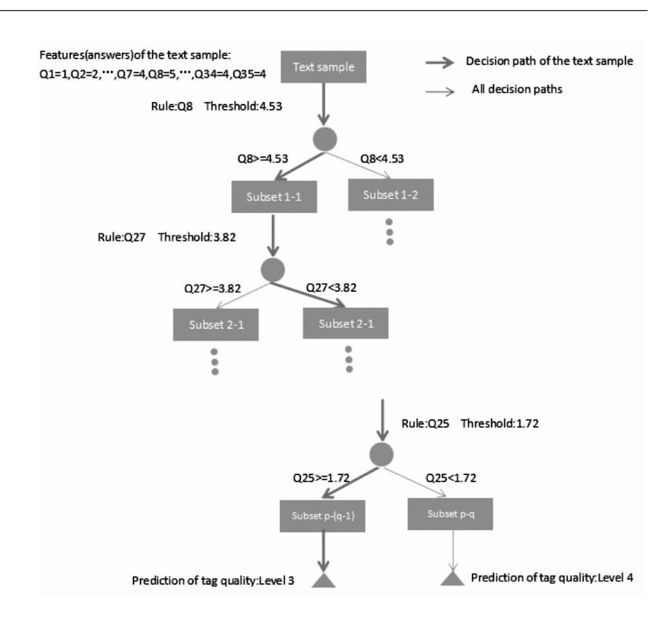


图 2 决策树模型及一个样本的预测过程示意图

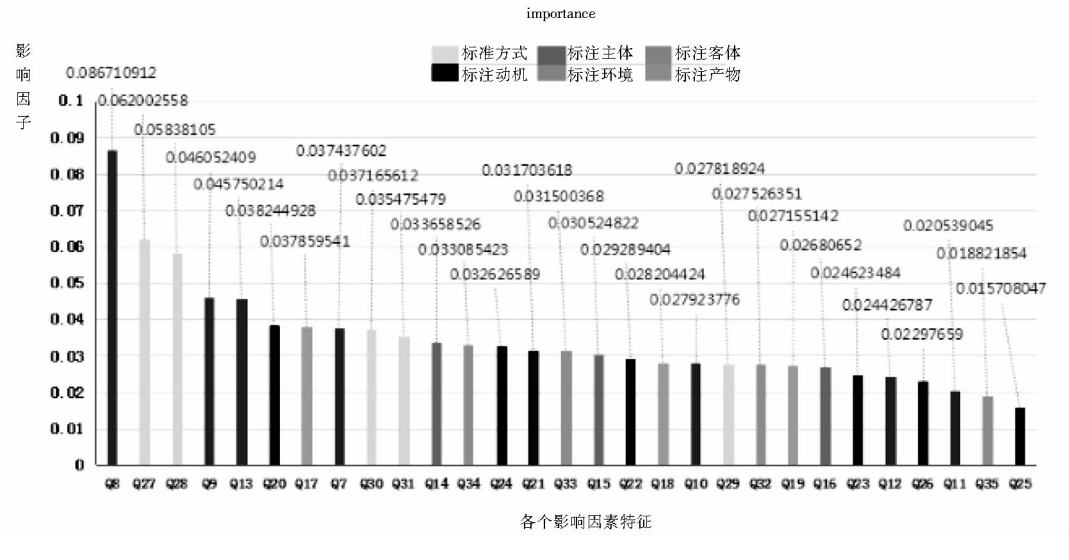


图 3 影响因素特征对标签质量的影响因子

如图 4 所示，6 个维度对于标签质量的影响权重按从高到低依次为标注主体、标注方式、标注动机、标注产物、标注客体和标注环境。

标注主体是对标签质量影响最大的因素。在该维度内，主体的知识结构和认知水平（Q8）、主体的标注频度（Q9）、主体的感知有用性（Q13）排在单因素中的第一位、第四位和第五位，对标签质量均影响显著。

标注方式的影响因子仅次于标注主体，排在第二位。其中，标签推荐（Q27）和规范标签提示（Q28）排在单因素中的第二、三位，也是对标签质量影响突出的因素。

标注动机的影响因子略低于标注方式，排在第三位。其中，揭示资源主题/分类/属性（Q20）的标注动机对标签质量的影响较大，排在单因素第六位。

除上述关键维度和关键因素外，标注产物、标注客体和标注环境 3 个维度对标签质量的影响因子均相对较低，但不乏其中单个因素对标签质量产生重要影响。

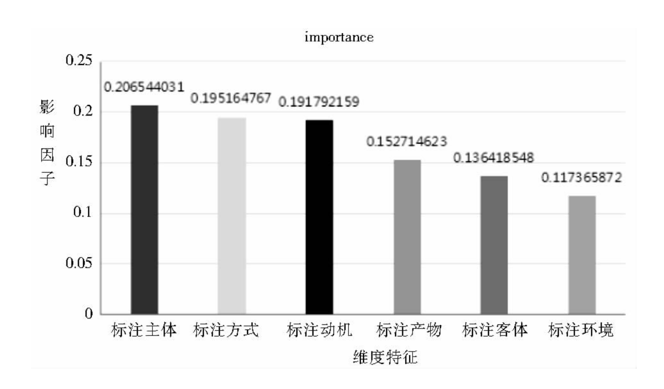


图 4　6 个维度特征对标签质量的影响因子

例如标注产物维度中的错拼(Q34)、标注客体维度中的资源类型(Q14)以及标注环境维度中的社会化标注系统的功能完整性(Q17)都是对标签质量有重要影响的因素。

理清社会化标注系统标签质量影响因素的权重，对针对性地提高社会化标注系统中标签质量意义重大，影响标签质量的关键维度具体如下：

(1)标注主体是影响标签质量的首要关键维度。其中，主体的知识结构和认知水平多由用户的受教育程度与生活背景决定，属于不易操控和改变的客观因素，故提高标签质量应注重从主体的标注频度和感知有用性入手。鉴于此，社会化标注系统应注重调动用户参与标注网络资源的积极性，通过提升用户的标注频率来积累标注经验，并使用户充分认识到标注高品质标签对于社会化标注系统网络资源分类、交流、共享的重要性，进而通过主体的主观能动性标注更多较高质量的标签。

(2)标注方式是影响标签质量的次要关键维度，其中标签推荐和规范标签提示是影响标签质量的重要因素。为提高标签质量，社会化标注系统可以采取如下措施：①设立标注指南，在标注区建立帮助文档说明标注的相关事项，方便用户掌握标注要点。如在标注区提供“标注过程详解”“标注注意事项”等。②设立规范标签推荐机制，采用分类词表、主题词表等受控词表或语义规范的高频标签列表等非受控词表引导的标签推荐机制，使得用户在标注资源时能在一定程度上优先选择系统推荐的规范标签；另外，也可引入“同义词环”等概念语义网络，在用户输入标签时向其推荐语义相近的规范标签，如当用户在输入“维他命”时，系统可出现“维生素”等规范词汇的推荐。③设置输入标签的长度限制和符号限制，如设定标签的最长字符数，规定标签不含标点符号，多个标签之间用分号隔开等，以此提高标签的精炼和准确性。④设置标签大小写、单复数、缩写词、结合词、简繁体等词汇控制机制，对于意义不明的缩略词应提醒用户进一步修改，避免缩略词引发的歧义(例如 CS 之于“Computer Science”和“游戏名称”)，使标签具有专指性。

(3)标注动机对于标签质量的影响排在第三位。其中，揭示资源主题/分类/属性的标注动机对标签质量的影响最大。因此，系统可设立标注维度的提示机制，鼓励并提示用户对资源的主题、分类、属性等维度展开重点标注，生成更多关于资源内容特征的高品质标签。

此外，提高标签质量也可从错拼、资源类型以及社会化标注系统的功能完整性等角度采取措施，具体如下：①社会化标注系统可建立标签纠错机制，当识别到用户标注的标签存在拼写错误时，系统会给用户提示或反馈修改建议，减少垃圾标签的产生；②调查数据表明，相对于富媒体类型资源，用户对文本资源更容易标注出高质量的标签，因而社会化标注系统应注重对资源类型进行划分，实现文本资源和富媒体资源的区别标注，使资源类型标注的分区设置在一定上实现标签质量的大致分级；③社会化标注系统应定期在平台中发布与用户体验和需求相关的调查问卷，依据用户需求完善标注、检索、导航、推荐、群组、分类等功能，优化标注流程，增加界面的设计亲和性和用户的使用便利性。

6　结论与展望

笔者从社会化标注系统中标签质量的影响因素出发，运用问卷调查法获得用户关于各个影响因素的看法和态度，并采用随机森林的机器学习算法建立标签质量影响因素的决策树模型，得出各个影响因素关于标签质量的影响因子。主要得出以下结论：标注主体是影响标签质量的首要关键维度，主体的知识结构和认知水平、标注频度及其感知有用性对标签质量的影响突出；标注方式是影响标签质量的次要关键维度，标签推荐和规范标签提示也是影响标签质量的重要因素。

与现有相关研究相比，本研究的重要价值和作用主要体现在 4 个方面：①关注的问题更普遍，聚焦于不同的社会化标注系统标签质量影响因素的共性，结论更具普适性；②梳理的影响因素更加全面，从标注主体、标注客体、标注环境、标注动机、标注方式、标注产物等角度综合考虑影响标签质量的各个因素，理论框架更加完善；③权重值的计算方法更科学，运用随机森林算法直接对调查数据进行客观训练得到权重，减少人为主观判断，求解过程更客观；④对策提出更具针对

性，从关键影响因素入手提出标签质量改进策略。

综上，本研究结果建立的标签质量影响因素预测模型，不仅解释性强，而且建立了影响因素特征和标签质量之间的显性表达式，提供了多属性权重值计算的新方法。权重值的确定不仅可以明晰当前标签低质量问题产生的原因，也可作为提高标签质量的参考依据，同时为标签质量评估体系中指标权重的确定提供了可靠依据。目前，标签质量评估大都基于标签的统计属性指标（如标签对应资源阅读次数、推荐数、评论数等）进行人工在线打分评估和自动化评估，但各个统计指标的权重却无从得知。本研究得到的标签质量影响因素的影响因子即可作为权重指标的参考，与统计指标体系相结合以更好地完善标签质量评估体系，这将是后续研究的方向之一。

参考文献：

[1] 郜杨芳，陈新国. 社会化标注系统中用户的标注行为及差异分析[J]. 图书馆，2017(10)：42－49，61.
[2] 熊回香，杨雪萍. 社会化标注系统中的个性化信息推荐研究[J]. 情报学报，2016，35(5)：549－560.
[3] 李旭晖，李媛媛，马费成. 我国图情领域社会化标签研究主要问题分析[J]. 图书情报工作，2018，62(16)：120－131.
[4] 章成志，李蕾. 社会化标签质量自动评估研究[J]. 现代图书情报技术，2015(10)：2－12.
[5] 章成志，赵华，李蕾，等. 中英文图片标签质量差异比较研究——以Flickr 为例[J]. 情报理论与实践，2018，41(4)：123－127.
[6] 黄如花，任其翔. WorldCat 热门标签的调查与分析[J]. 图书与情报，2012(5)：7－10.
[7] 吴方枝. Flickr 网站用户标签的质量控制对策[J]. 图书馆学研究，2012(11)：26－28.
[8] SOGOL N，ARASH B，CHEN D. An improved collaborative recommendation system by integration of social tagging data[EB/OL]. [2018－08－01]. https://doi.org/10.1007/978－3－319－14379－8_7.
[9] HALL C E，ZARRO M A. What do you call it?：a comparison of library-created and user-created tags[C]//Proceedings of the11th annual international ACM/IEEE joint conference on digital libraries. New York：ACM，2011：53－56.
[10] GUY M，TONKIN E. Folksonomies：tidying up tags? [EB/OL]. [2018－08－02]. http://www.dlib.org/dlib/january06/guy/01guy.html#1.
[11] FlOECK F，PUTZKE J，STEINFELS S，et al. Imitation and quality of tags in social bookmarking systems[C]//Advances in intelligent and soft computing. Berlin：Springer-verlag，2010：75－91.
[12] SEN S，HARPER F M，LAPITZ A，et al. The quest for quality tags[EB/OL]. [2018－08－01]. http://www.doc88.com/p-3724514050617.html.
[13] 吴克文，朱庆华，赵宇翔，等. 社会化标注系统中标签检索质量模拟研究[J]. 情报学报，2011，30(1)：29－36.
[14] 罗琳，杨洋. 社会化标注系统中用户标签使用行为影响因素研究[J]. 图书情报知识，2018(3)：85－94.
[15] 高兵，孙琳，谢彪，等. 权重概率主成分分析模型的建立及应用研究[J]. 中国卫生统计，2018，35(6)：802－805.
[16] 徐少成，李东喜. 基于随机森林的加权特征选择算法[J]. 统计与决策，2018，34(18)：25－28.
[17] SCOTT A G，BERNARDO A H. Usage patterns of collaborative tagging systems[J]. Journal of information science，2006，32(2)：198－208.
[18] 刘竞妍，张可，王桂华. 综合评价中数据标准化方法比较研究[J]. 数字技术与应用，2018，36(6)：84－85.

作者贡献说明：

张云中：确定论文选题，提出整体研究思路和框架，修改定稿；
秦艺源：负责数据处理和分析，论文起草及修改。

Research on Influencing Factors of Tag Quality in Social Tagging System：Based on Random Forest

Zhang Yunzhong Qin Yiyuan
Department of Library，Information and Archives，Shanghai University，Shanghai 200444

Abstract：[**Purpose/significance**] Tag quality is often related to users' experience of classification，query，browsing，acquisition of online resources in social tagging system. Identifying key influencing factors of tag quality can optimize the core functions of resources organization of STS. [**Method/process**] Based on tags，we provided the influencing factors model of tag quality from six perspectives，which covered tagging subject，tagging object，tagging environment，tagging motivation，tagging methods and tagging products. The study attempted to explore the key influencing factors of tag quality by questionnaire，and established the decision tree model of influencing factors of tag quality based on Random Forest. [**Result/conclusion**] Tagging subject is the primary key dimension affecting tag quality. And the impact of the subject's knowledge structure and cognitive level，the subject's tagging frequency，and the subject's perceived usefulness are prominent. Tagging methods are the secondary one，and tag recommendation and standard tag tips are main influencing factors.

Keywords： social tagging system tag quality machine learning random forest